Reliability of Clinician Rated Physical Effort Determination During Functional Capacity Evaluation in Patients with Chronic Musculoskeletal Pain

Trippolini, M. A.; Dijkstra, P. U.; Jansen, B.; Oesch, P.; Geertzen, J. H. B.; Reneman, M. F.

Published in:
Journal of Occupational Rehabilitation

DOI:
10.1007/s10926-013-9470-9

Publication date:
2014

J Occup Rehabil (2014) 24:361–369
DOI 10.1007/s10926-013-9470-9

Reliability of Clinician Rated Physical Effort Determination During Functional Capacity Evaluation in Patients with Chronic Musculoskeletal Pain

M. A. Trippolini · P. U. Dijkstra · B. Jansen · P. Oesch · J. H. B. Geertzen · M. F. Reneman

Published online: 22 August 2013

Abstract *Introduction* Functional capacity evaluation (FCE) can be used to make clinical decisions regarding fitness-for-work. During FCE the evaluator attempts to assess the amount of physical effort of the patient. The aim of this study is to analyze the reliability of physical effort determination using observational criteria during FCE. *Methods* Twenty-one raters assessed physical effort in 18 video-recorded FCE tests independently on two occasions, 10 months apart. Physical effort was rated on a categorical four-point physical effort determination scale (P_{ED}) based on the Isernhagen criteria, and a dichotomous submaximal effort determination scale (S_{ED}). Cohen's Kappa, squared weighted Kappa and % agreement were calculated. *Results* Kappa values for intra-rater reliability of P_{ED} and S_{ED} for all FCE tests were 0.49 and 0.68 respectively. Kappa values for inter-rater reliability of P_{ED} for all FCE tests in the first and the second session were 0.51, and 0.72, and for S_{ED} Kappa values were 0.68 and 0.77 respectively. The inter-rater reliability of P_{ED} ranged from $\kappa = 0.02$ to $\kappa = 0.99$ between FCE tests. Acceptable reliability scores ($\kappa > 0.60$, agreement $\geq$80 %) for each FCE test were observed in 38 % of scores for P_{ED} and 67 % for S_{ED}. On average material handling tests had a higher reliability than postural tolerance and ambulatory tests. *Conclusion* Dichotomous ratings of submaximal effort are more reliable than categorical criteria to determine physical effort in FCE tests. Regular education and training may improve the reliability of observational criteria for effort determination.

Keywords Rehabilitation · Pain · Disability evaluation · Lifting · Work capacity evaluation

M. A. Trippolini (✉) · B. Jansen
Department of Work Rehabilitation, Rehaklinik Bellikon, Suva Care, 5454 Bellikon, Switzerland
e-mail: maurizio.trippolini@rehabellikon.ch

M. A. Trippolini · P. U. Dijkstra · J. H. B. Geertzen · M. F. Reneman
Department of Rehabilitation Medicine, Center for Rehabilitation, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

P. U. Dijkstra
Department of Oral and Maxillofacial Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

P. Oesch
Research Department, Rehabilitation Centre Valens, Valens, Switzerland

P. Oesch
Department of Rheumatology, Rehabilitation Centre Valens, Valens, Switzerland

Abbreviations

FCE	Functional capacity evaluations
P_{ED}	Physical effort determination by categorical observational criteria
S_{ED}	Submaximal effort determination by dichotomous rating
CMP	Chronic nonspecific musculoskeletal pain

Introduction

Individuals suffering from chronic nonspecific musculoskeletal pain (CMP) such as back and neck pain are often restricted in performing activities of daily living and work [1, 2]. The financial burden of CMP on society arises mainly due to indirect costs because of temporary or permanent work disability. Work disability due to CMP may

be associated with reduced activity levels and work performance [3, 4]. Functional capacity evaluation (FCE) in addition to self-reported measures have been recommended for a comprehensive assessment of physical work performance for persons with CMP [5–8].

Functional capacity evaluation employs physical performance tests such as lifting, postural tolerance tests, repetitive movements, and ambulation to assess work-related functioning [9]. Discrepancies in FCE outcomes and the physical workload of a patient may be addressed in rehabilitation to restore this imbalance [10–12]. Moreover, FCEs are used to evaluate the effects of rehabilitation and determine fitness-for-work, and as such FCEs may facilitate the return-to-work process or prelude case closure [13–17].

To determine physical capacity during the FCE the patient must perform to his or her maximum level of physical ability. The level of physical effort during FCE is estimated by the evaluator, based on observational criteria during material and non-material handling tests [9, 18]. Submaximal effort is assumed when a person stops a FCE test before the criteria indicative of maximal effort are observed. Because clinical decision-making is based on the results of FCE, sound clinimetric properties of observational criteria are required to determine physical effort. Acceptable reliability of physical effort determination FCE tests such as lifting has been reported [19, 20]. However, the reliability of non-material handling tests such as kneeling and forward bending has rarely been studied [21–25]. Moreover, most studies on lifting tests were performed by FCE experts, which limits the generalizability and applicability of the study results among less experienced raters [25–27].

The aim of this study was to determine the intra- and inter-rater reliability of physical effort determination of FCE tests in patients with CMP. A second aim was to investigate whether an increase in rater experience would alter the reliability of physical effort determination.

Methods

Procedures, Patients and Video Sequences

Video tape-recordings were taken during FCEs, performed in a work rehabilitation setting. FCE tests were performed according to the Isernhagen test procedure, which claims to measure a person's physical capacity to safely engage in work-related activity [28]. Four patients (3 with non-specific low back pain and 1 with non-specific neck pain, mean age 35.5 years, range 21–49 years) were recruited based on convenience. All patients were instructed how to perform the test, and that they were expected to perform maximally. Testing could be terminated for four reasons: the participant stopped because of, for example, pain; the observer deemed testing to have become over safe maximum based on criteria for effort determination (Appendices 1, 2); heart rate exceeded 85 % of the age-related maximum (220 minus age of participant); or a predefined time limit was reached. All patients gave written consent to be video-recorded. Eighteen videos from 11 FCE tests with a total duration of 28 min were selected. The videos were mute recorded. For each test information was provided on a standardized form regarding heart rate at the beginning and end of the test, and weight lifted in kilograms (for material handling tests) or duration (for static posture, or walking, stair climbing).

Raters

A convenience sample of 21 physiotherapists (11 female, 10 male) from Bellikon rehabilitation clinic (Switzerland) served as a representative sample of raters. Nineteen had attended the official 2-day FCE training course provided by the Swiss Rehabilitation Association [18]. Prior to the study all had performed at least ten 1-day FCEs in the previous year [median 30, interquartile range (IQR) 20–33] and had a minimum of 1 year work experience in work rehabilitation (median 3, IQR 2–3), and a minimum professional practice experience of 1 year (median 5 years, IQR 3–12.5).

Physical Effort Determination During FCE Tests

The 18 videos were shown in a classroom to all the raters at the same time. Prior to the showing the raters were instructed about the procedure of the rating. The ratings of physical effort were filled in a standardized form with a pencil. The videos consisted of 18 tests. When a test was finished and all participants had rated that test, then the next test was shown. Raters were not allowed rewind the video or to stop a video while a test was shown. Each video was shown once per session. Raters were blinded each other's ratings. Each video was rated according to observational criteria indicative of physical effort for material handling tests as "light to moderate", "heavy" or "maximal" (Appendix 1). Observational criteria for postural tolerance tests and ambulation tests were rated on a scale from "No or slight functional problem/limitation", "some functional problem/limitation" to "substantial functional problem/limitation" (Appendix 2). This categorical scale was termed physical effort determination (P_{ED}) scale. If a test was performed unsafely it was classified as "over safe maximum", when observed performance exceeded the maximum observational criteria for physical effort level during work-related tasks (Appendices 1, 2). Tests were scored as "not classifiable" when the patient interrupted

the FCE test at the very start or the observed effort was not clearly interpretable to the raters and no conclusions could be drawn. Submaximal effort was assumed when a patient stopped a material or non-material handling test before the FCE rater observed sufficient criteria indicative of maximal weight, or significant functional problems/limitation as described in Appendices 1 and 2. This dichotomous scale was termed submaximal effort determination (S_{ED}).

Maximal effort was defined as the highest safe ability of a person during a FCE test [9]. An FCE was considered safe when no formal complaints of injury or serious adverse effects were filed by the patients, and when increased symptoms returned to or below their pre-FCE level [29].

The observers rated each video twice, in September 2010 (session 1) and in July 2011 (session 2). Between these sessions each rater performed approximately 30 short FCEs (material handling tests only), as part of the regular clinical procedure of a work rehabilitation program. All raters attended both sessions. Data extraction into the database was performed by an individual who was not involved in the data analysis.

Both patients and raters agreed that their data would be used either for the scope of research or education. Because this study was part a regular educational video based training, no ethical approval was required. However, this study was part of a research project approved by the Medical Ethics Committee of Canton Aargau, Switzerland (EK AG 2010/055) [30].

Data Analysis

Intra-rater reliability was assessed by comparing the scores from the first rating session with the scores from the second session for each rater. Inter-rater reliability was assessed twice: by comparing the scores between all the raters in session 1 and 2. Category 5 "not classifiable" was excluded from the analyses. Inter-rater and intra-rater reliability was calculated using Cohen's Kappa values for dichotomous data, and squared weighted Kappa values for categorical data and percentages of agreement. A percentage of agreement of 80 % or more was judged as acceptable. If agreement was ≥80 % and Kappa was $\kappa > 0.60$ then reliability values were considered as acceptable [31]. AGREE (Agree, Version 7.002) was used to analyze Kappa for multiple observer categories [32] and the ONLINE KAPPA CALCULATOR was used for multiple raters [33]. All other analyses were performed using SPSS (Statistical Package for Social Sciences, Version 20, 2011).

Results

Intra-rater Reliability of Physical Effort Determination for all FCE Tests

Excluding category 5 "not classifiable" resulted in 325 ratings for the categorical scale for physical effort determination (P_{ED}) (Table 1) and 376 ratings were performed for the dichotomous scale for submaximal effort (S_{ED}) (Table 2).

Reliability of Physical Effort Determination (P_{ED})

The intra-rater reliability of P_{ED} for all FCE tests in both sessions together was $\kappa = 0.49$ (95 % CI 0.22–0.75). The inter-rater agreement of P_{ED} for all FCE tests increased from 73 % (session 1) to 85 % (session 2). Kappa values as a measure of inter-rater reliability of P_{ED} for all FCE tests increased from session 1 (0.51; 95 % CI 0.23–0.80) to session 2 (0.72; 95 % CI 0.49–0.94). Mean Kappa values for inter-rater reliability of P_{ED} increased from session 1 to 2 for material handling (0.17), postural tolerance (0.21) and ambulation (0.03) (Table 3). Mean agreement values of material handling, postural tolerance and ambulation tests ranged from 54 to 75 % for inter- and intra-rater reliability (Table 3).

Table 1 Cross tabulation of the categorical ratings for physical effort determination (P_{ED}) in session 1 and 2

		Description	Session 2					Total
	Category[a]		1	2	3	4	5	
Session 1	1	Light to medium effort	156	32	2	1	4	195
	2	Heavy effort	40	70	5	1	5	121
	3	Maximum effort	2	8	5	0	8	23
	4	Over safe maximum	0	3	0	0	0	3
	5	Not classifiable[b]	7	2	0	0	27	36
Total			205	115	12	2	44	378

[a] Categories 1–5 are described in Appendices 1 and 2; [b]Category 5 "not classifiable" was excluded from the analyses

Table 2 Cross tabulation of the categorical ratings for submaximal effort determination scale (S_{ED}) in session 1 and 2

	Category[b]	Session 2		Total
		Criteria for maximal physical effort observed[a]		
		Yes	No	
Session 1	Yes	241	27	268
	No	23	85	108
Total		264	112	376

[a] Yes = observed effort was assumed to be indicative for maximal effort as described in Appendices 1 and 2 when patient performed the material or non-material handling test. [b] No = Submaximal effort was assumed when a patient stopped a material or non-material handling test *before* the FCE rater observed sufficient criteria indicative of maximal weight, or significant functional problems/limitation as described in Appendices 1 and 2

Table 3 Inter- and intra-rater reliability for each FCE test

Category	Test (n)	Physical effort determination scale (P_{ED})[a]						Submaximal effort scale (S_{ED})[b]					
			Inter		Inter	Intra			Inter		Inter	Intra	
		Session 1	Session 1	Session 2	Session 2	Session 1–2		Session 1	Session 1	Session 2	Session 2	Session 1–2	
		%	κ	%	κ	%	κ	%	κ	%	κ	%	κ
M	One-handed carrying (4)	68	0.57	*80*	*0.74*	71	0.54	75	0.49	75	0.49	76	0.29
M	Lifting floor to waist (4)	58	0.43	73	*0.64*	67	0.47	*85*	*0.70*	*88*	*0.76*	*85*	0.05
M	Two-handed horizontal lift (2)	50	0.34	47	0.29	66	0.34	*95*	*0.90*	*100*	*1.00*	*100*	*1.00*
M	Lifting waist to overhead (2)	66	0.55	*91*	*0.88*	*81*	0.60	*100*	*1.00*	*100*	*1.00*	*100*	*1.00*
Mean		61	0.47	73	0.64	71	0.49	89	0.77	91	0.81	90	0.59
P	Kneeling (1)	*80*	*0.73*	*90*	*0.99*	*84*	−0.08	68	0.35	*100*	*1.00*	*100*	*1.00*
P	Forward bend sitting (1)	44	0.25	33	0.11	55	NA	68	0.35	*90*	*0.81*	76	−0.08
P	Overhead working (1)	42	0.22	79	*0.72*	50	0.35	74	0.49	*90*	*0.81*	*80*	−0.08
Mean		55	0.40	67	0.61	63	0.14	70	0.40	93	0.87	85	0.28
A	Stair climbing (1)[c]	62	0.49	*100*	*1.00*	76	0.00	*90*	*0.80*	*100*	*1.00*	*100*	*1.00*
A	Stair climbing (1)[d]	27	0.02	0	−0.33	74	NA	*100*	*1.00*	*100*	*1.00*	*100*	*1.00*
A	Walking (1)	73	*0.64*	68	0.57	75	0.14	56	0.12	57	0.14	*90*	*0.76*
Mean		54	0.38	56	0.41	75	0.07	82	0.64	86	0.71	97	0.92

Inter: inter-rater reliability; intra: intra-rater reliability; %: percentage agreement; κ: Cohen's Kappa values for dichotomous, Squared weighted Kappa for categorical data; [a] observational criteria for determination of physical effort during material and non-material handling tests (see Appendices 1, 2); [b] submaximal effort was assumed, when a participant stopped a material or non-material handling tests before the FCE rater observed sufficient observational criteria indicative of maximal effort; M: material handling tests; P: postural tolerance tests; A: ambulation tests; (n): number of videos; [c] short video length until patient stops; [d] full video length of the test 10 × 10 stairs up and down; NA: not applicable, due to lack of cell filling. Italicised values criteria for acceptable reliability (agreement ≥80 %, $\kappa > 0.60$)

Reliability of Submaximal Effort Determination (S_{ED})

For S_{ED} the intra-rater reliability for all FCE tests in both sessions together was $\kappa = 0.68$ (95 % CI 0.60–0.76). Kappa values as a measure of inter-rater reliability of S_{ED} for all FCE tests increased from session 1 (0.68; 95 % CI 0.60–0.76) to session 2 (0.77; 95 % CI 0.70–0.84). Mean Kappa values for inter-rater reliability of S_{ED} increased

from session 1 to 2 for material handling (0.04), postural tolerance (0.47) and ambulation (0.07) (Table 3). Mean agreement values of material handling, postural tolerance and ambulation tests ranged from 70 to 97 % for inter- and intra-rater reliability (Table 3).

Comparison Reliability of P_{ED} and S_{ED}

In 6 out of 10 tests inter-rater agreement and Kappa values for the P_{ED} were equal or increased from session 1 to session 2. For S_{ED} inter-rater agreement and Kappa values were similar or increased for all 10 tests. The general reliability of S_{ED} was higher than that of P_{ED}. The inter-rater reliability (% agreement) of S_{ED} was higher in 8 tests (out of 10) for session 1, and in 8 tests (out of 10) for session 2 than that of P_{ED}. The inter-rater reliability (Kappa) of S_{ED} was higher in 7 tests (out of 10) for session 1, and in 8 tests (out of 10) for session 2 than that of P_{ED}. For intra-rater reliability (% agreement/Kappa) S_{ED} was higher than P_{ED} in 10 out of 10 and 5 out of 10 tests respectively.

When applying cut-off scores for acceptable reliability (agreement levels $\geq$80 %, $\kappa > 0.60$), 46 % (55 out of 120) of the reliability values fulfilled this criterion (see italicised values in Table 3).

Discussion

When applying cut-off scores of agreement $\geq$80 %, $\kappa > 0.60$, the overall reliability of P_{ED} and S_{ED} was acceptable for less than half (46 %) of all FCE observations. For S_{ED} reliability was acceptable in the majority (67 %) of the FCE tests. However, the reliability of the P_{ED} was acceptable in only 38 % of tests. Inter- and intra-rater reliability between each FCE test varied considerably. The increase in mean reliability scores from session 1 to session 2 was on average higher in the P_{ED} than in the S_{ED}.

S_{ED} during FCE tests can be reliably detected in the majority of cases. However the results of this study are disappointing, as raters reached the required reliability cut-off values for both the P_{ED} and S_{ED} in less than half of the observations. This finding has clinical relevance for four reasons. First: some FCEs claim to support fitness-for-work determination with an extrapolation of FCE results to job demands [14, 34]. The job demands and their frequencies during a working day (occasional, 1–33 %; frequent, 34–66 %; constant 67–100 %) are matched to P_{ED} "maximum", "heavy" and "light to moderate". Good reliability of P_{ED} is needed to enable adequate matching between FCE performance and work demands. Second: FCEs have been reported to accurately describe physical capacity only if a person exerts "maximal" voluntary effort [23, 35]. Good reliability of determination of effort is a prerequisite for such a clinical interpretation. Third: FCE reports are used by third parties to inform on the progress of insurance claims. Some interpret submaximal physical effort as 'unmotivated'. The debate over whether this interpretation is valid is beyond the scope of this paper, but it highlights the relevance of the psychometric properties of this determination. Fourth: whether the FCE score represents maximal or submaximal capacity, and the reasons for performing submaximally, are relevant for designing individualized vocational rehabilitation aimed at improvement of functional capacity.

Compared to three previous reliability studies on material handling tests, our values are clearly lower [22, 23, 26]. In some of these previous studies with high reliability values two-point scales for determination of physical effort were used, which increases the a priori probability for agreement compared to a multiple item scale as in our study. In our study agreement on the dichotomous scale (submaximal effort determination) was substantially higher too. Moreover the results show on average an increase in the agreement and reliability rating on both the P_{ED} and S_{ED} scales when administered 10 months apart, indicating a "learning" effect. Our data support the assumption that postural tolerance tests may be difficult to rate using the FCE observational methods, but that experience can substantially improve reliability. The average agreement and Kappa values for the inter-rater reliability of P_{ED} increased by 0.40 during the 10-month period. This may be partly attributed to experience. The raters participating in this study used 1-day FCEs for the standard assessment of most in-patients. In addition they received one-to-one supervision from an FCE expert once a year, and their superiors supervised each FCE report as part of regular quality control. Based on the observation in this study that experience and basic training increased reliability scores, we suggest that novice raters using the observational criteria are supervised more intensively than in our study. To what extent observational criteria for effort determination can be improved by additional training remains unknown.

The only slight increase in the agreement and reliability of S_{ED} might be due to the high scores obtained in the first observation session. When tests were grouped according to type of task the reliability of the physical effort determination scale was generally lower when applied to postural tolerance tests, such as overhead working and kneeling, than when used with material handling tests. This is consistent with results from studies reporting on forward bend, standing and crouching [25, 35, 36]. Moreover observational criteria seem to be less reliable when applied to ambulation tests such as walking and stair climbing compared to material handling tests [25, 36]. However, the

results may be influenced by the fact that postural tolerance tests were not part of the regular 1-day FCE utilized in most in-patients, but were only used when indicated. Thus, raters collected more test-experience with the observation of material handling tests than with postural tolerance. Other possible reasons for the lower reliability of the postural tolerance and ambulation tests could be the ceiling effect due to the predefined maximal time limit of the test or the muscular use at submaximal rates. It is theoretically infeasible to judge maximum effort level when submaximal muscular effort is requested e.g. in the overhead work test, the duration of 5 min is not the requested maximum performance, but a time limit. The results of this study underscore this problem. We suggest that observational criteria of physical effort in postural tolerance and ambulation tests need further refinement. To our knowledge no study has been conducted to determine the validity of observational criteria for postural tolerance and ambulation tests in FCE.

In two videos in which a patient performed the one-handed carrying test, ratings showed low agreement. After rating, we discussed these two videos with the raters and asked them where the difficulty lay. Almost half of the raters responded that these were debatable videos due to the pain behavior of the patient. The maximum performance of a patient is determined by the individuals' ability, motivation, and other psychosocial factors [37, 38]. However, physical effort determination cannot be used interchangeably with non-organic signs described by Waddell et al., despite some important overlap of the two measurement methods [38]. It has been questioned whether lay persons and health care providers can accurately classify effort during a lifting task performed by actors [39]. Similarly to our results this underscores the challenge of determining effort using a categorical rating scale.

Strengths and Weaknesses of the Study

The strengths of the study were that the inter- and intra-rater reliability measures were based on the results of a large sample of raters, and multiple observations on patient videos. Compared to most other studies on the reliability of P_{ED}, additionally to the material handling tests, we included postural tolerance and ambulatory tests. Furthermore this is to our knowledge the first study on the reliability of observational criteria used in FCE tests based on two ratings taken within a period of 10 months, excluding the risk of recall bias. We used 18 videos instead of real patients to test the reliability of the observers. The results may therefore only partly reflect a FCE performed live with the patient. One may argue that several clinical parameters may not have been visible on video tape, such as respiration, and that the raters did not benefit from three-dimensional vision. Observing videos without sound and communication is relevantly different from a clinical setting. In clinical practice FCE raters observe the same patient at different levels of effort when performing the same FCE test. This might facilitate comparison of their own ratings with their previous observations. Studies should be performed to analyze whether the availability of additional information would have changed the results. This study was performed with a sample of four patients. We might therefore not have seen all types of movement patterns of patients with back pain. Because the study was designed to measure the reliability of the raters observing the performance rather than the reliability of that performance, this may have been adequate. The Kappa statistic has an advantage over percentage of agreement because it corrects for chance [31]. In some tests high agreement between raters was observed and Kappa values were in some cases extremely low. This phenomenon may occur when the variation in row and column totals is low [40]. Furthermore it may be debatable if the cut-off score for Kappa values of $\kappa > 0.60$ for acceptable reliability used in our study is enough rigorous when one has to make decisions at the individual patient level [41]. The results should therefore be interpreted accordingly. Category 5, "not classifiable", was excluded from the analysis for two reasons. First "not classifiable" relates to another dimension than those categories related to effort. Therefore it cannot be analyzed in the effort domain. Secondly, only a few ratings were "not classifiable", indicating its minor influence.

Future Studies

Although there have been some advances in the study of reliability of physical effort determination, major gaps remain: for example, what are valid and practical reference standards for determining maximal physical effort during FCE tests? While some experimental studies measuring muscle activity measurements such as surface EMG, superimposed electrical stimulation, and lactate concentration have been performed, they lack practicality for clinical use [42, 43]. How should evidence-based cut-off scores of reliability be defined that are useful for the various purposes of FCE? Future studies should address these unresolved questions and promote the development of a reliable tool for the determination of physical effort, above all for postural tolerance tests.

Conclusions

The reliability of observing physical effort varied substantially between FCE tests, ranging from unacceptable to good. The dichotomous rating of sub-maximal effort was more reliable than the categorical rating for physical effort determination. However, with both rating scales acceptable reliability values were reached on average only in every second observation, which limits their utility for clinical decision-making. Regular education and training may improve the reliability of observational criteria for effort determination. Further research is needed to develop reliable observation scales.

Acknowledgments The authors thank the physiotherapists of the Department of Work Rehabilitation, Rehaklinik Bellikon who participated in this study. We also thank Doug Gross and Dee Delay for the fruitful discussions on the criteria for physical effort determination. Part of the study was funded by the Swiss Accident Insurance Fund, SUVA (Schweizerische Unfallversicherungsanstalt).

Conflict of interest We certify that no party having a direct interest in the results of the research supporting this article has or will confer a benefit on us or on any organization with which we are associated AND, if applicable, we certify that all financial and material support for this research (e.g., NIH or NHS grants) and work are clearly identified in the manuscript.

Appendix 1

See Table 4.

Table 4 Observational criteria for determination of physical effort during material handling tests

Criteria	Light to moderate	Heavy	Maximum
Muscle recruitment			
Prime movers	Normal recruitment	Bulging	Bulging
Accessory muscles	No or only slight muscle recruitment	Distinct recruitment	Bulging
Base of support	Natural stance	Distinctly increased	Very wide base
Posture	No or only slight counterbalance in extension	Distinctly increased counterbalance	Substantial counterbalance
Heart rate and respiration	No or minimal increases in heart rate and respiration	Distinct increases in heart rate and respiration	Substantial increases in heart rate and respiration
Control and safety	Smooth movements	Increasingly controlled movement; might begin to use momentum; execution with difficulty but not yet at the limit	Still safe but unable to maintain control with the addition of any more weight
Pace	Moderate/comfortable pace	Distinctly slower; very deliberate movements	Very slow (an increased pace would affect stability and control)

The level of physical effort during material handling tests was determined on the basis of observational criteria indicative of light to moderate, heavy, or maximal weight load [9, 18, 44]. Maximal effort was assumed when, on the basis of the expertise of the functional capacity evaluation (FCE) rater, sufficient criteria indicative of safe maximal weight were observed. Submaximal effort was assumed when a participant stopped a material handling test before the FCE assessor observed sufficient criteria indicative of maximal weight. Appendix 1 is used with permission from Verein IG Ergonomie, Swiss Association of Rehabilitation

Appendix 2

See Table 5.

Table 5 Observational criteria for determination of physical effort during non-material handling tests

Criteria	No or slight functional problem/ limitation	Some functional problem/ limitation	Substantial functional problem/ limitation
Posture	Maintains normal posture, or slight deviation in posture[a]	Some deviation from normal posture[a], occasional change of position	Substantial deviation from normal posture[a], substantial unrest (frequent change of posture position)
Movement pattern	Normal movement pattern, slight deviation from normal[a], smooth movements or slight muscle stiffness, normal to slightly slower performance	Some deviation from the normal movement pattern[a], tense movements, markedly slower performance	Substantial deviation from the normal movement pattern[a], very tense movements, very slow performance
Muscle recruitment	Normal recruitment of prime movers only, or minimal recruitment of accessory and stabilizing muscles of the trunk, neck or joints stabilizers	Some recruitment of accessory and stabilizing muscles of the trunk, neck or joints stabilizers	Pronounced recruitment of accessory and stabilizing muscles of the trunk, neck or joints
Reaction of the autonomic nervous system	Minimal increase in heart rate	Moderate increase in heart rate and respiration	Substantial increase in heart rate, respiration rate and significant sweating

The level of physical effort during non-material handling tests was determined on the basis of observational criteria indicative of no or slight limitation/problem, some functional limitation/problem, or significant limitation/problem [28]. Maximal effort was assumed when, on the basis of the expertise of the functional capacity evaluation (FCE) rater, sufficient criteria indicative of substantial functional problem/limitation were observed. Submaximal effort was assumed when a participant stopped a non-material handling test before the FCE rater observed sufficient criteria of substantial functional problem/limitation. Appendix 2 is used with permission from Verein IG Ergonomie, Swiss Association of Rehabilitation

[a] Asymmetry (unequal loading) or deviation from neutral

References

1. Gerdle B, Bjork J, Henriksson C, Bengtsson A. Prevalence of current and chronic pain and their influences upon work and healthcare-seeking: a population study. J Rheumatol. 2004;31:1399–406.
2. Bevan S, Quadrello T, McGee R, Mahdon M, Vavrovsky A, Barham L. Fit for work? Musculoskeletal disorders in the European workforce. UK: The Work Foundation; 2009.
3. Lambeek LC, van Mechelen W, Knol DL, Loisel P, Anema JR. Randomised controlled trial of integrated care to reduce disability from chronic low back pain in working and private life. BMJ. 2010;340:1035.
4. Alschuler KN, Theisen-Goodvich ME, Haig AJ, Geisser ME. A comparison of the relationship between depression, perceived disability, and physical performance in persons with chronic pain. Eur J Pain. 2008;12:757–64.
5. Schiphorst Preuper HR, Reneman MF, Boonstra AM, Dijkstra PU, Versteegen GJ, Geertzen JH. The relationship between psychosocial distress and disability assessed by the Symptom Checklist-90-Revised and Roland Morris Disability Questionnaire in patients with chronic low back pain. Spine J. 2007;7:525–30.
6. Smeets RJ, van Geel AC, Kester AD, Knottnerus JA. Physical capacity tasks in chronic low back pain: what is the contributing role of cardiovascular capacity, pain and psychological factors? Disabil Rehabil. 2007;29:577–86.
7. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2005;113:9–19.
8. Wittink H. Functional capacity testing in patients with chronic pain. Clin J Pain. 2005;21:197–9.
9. Isernhagen SJ. Functional capacity evaluation: rational, procedure, utility of the kinesiophysical approach. J Occup Rehabil. 1992;2:157–68.
10. Henchoz Y, de Goumoens P, Norberg M, Paillex R, So AK. Role of physical exercise in low back pain rehabilitation: a randomized controlled trial of a three-month exercise program in patients who have completed multidisciplinary rehabilitation. Spine (Phila Pa 1976). 2010;35:1192–9.
11. Isernhagen SJ, Galper JS. General testing principles for functional capacity evaluations. In: Genovese E, Galper JS, editors. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago, IL: American Medical Association; 2009. p. 41–52.
12. Kool JP, Oesch PR, Bachmann S, Knuesel O, Dierkes JG, Russo M, et al. Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: a randomized controlled trial. Arch Phys Med Rehabil. 2005;86:857–64.
13. Durand MJ, Brassard B, Hong QN, Lemaire J, Loisel P. Responsiveness of the physical work performance evaluation, a functional capacity evaluation, in patients with low back pain. J Occup Rehabil. 2008;18:58–67.

14. Oesch PR, Kool JP, Bachmann S, Devereux J. The influence of a functional capacity evaluation on fitness for work certificates in patients with non-specific chronic low back pain. Work. 2006;26:259–71.
15. Wind H, Gouttebarge V, Kuijer PP, Sluiter JK, Frings-Dresen MH. Complementary value of functional capacity evaluation for physicians in assessing the physical work ability of workers with musculoskeletal disorders. Int Arch Occup Environ Health. 2009;82:435–43.
16. Kuijer PP, Gouttebarge V, Brouwer S, Reneman MF, Frings-Dresen MH. Are performance-based measures predictive of work participation in patients with musculoskeletal disorders? A systematic review. Int Arch Occup Environ Health. 2011;85:109–23.
17. Genovese E, Isernhagen SJ. Approach to requesting a functional evaluation. In: Genovese E, Galper JS, editors. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago, IL: American Medical Association; 2009. p. 19–40.
18. Denier-Bont F, Fischer V, Oesch P, Oliveri M. Functional capacity evaluation: course manual. Bellikon: Verein IG Ergonomie, Swiss Association of Rehabilitation; 2007.
19. Innes E. Handgrip strength testing: a review of the literature. Aust Occup Ther J. 1999;46:120–40.
20. Gouttebarge V, Wind H, Kuijer PP, Frings-Dresen MH. Reliability and validity of functional capacity evaluation methods: a systematic review with reference to Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen work system. Int Arch Occup Environ Health. 2004;77:527–37.
21. Reneman MF, Fokkens AS, Dijkstra PU, Geertzen JH, Groothoff JW. Testing lifting capacity: validity of determining effort level by means of observation. Spine (Phila Pa 1976). 2005;30:E40–6.
22. Smith RL. Therapists' ability to identify safe maximum lifting in low back pain patients during functional capacity evaluation. J Orthop Sports Phys Ther. 1994;19:277–81.
23. Gardener L. Reliability of occupational therapists in determining safe, maximal lifting capacity. Aust Occup Ther J. 1999;46:110–9.
24. Jay MA, Lamb JM, Watson RL, Young IA, Fearon FJ, Alday JM, et al. Sensitivity and specificity of the indicators of sincere effort of the EPIC lift capacity test on a previously injured population. Spine (Phila Pa 1976). 2000;25:1405–12.
25. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN. Test–retest reliability of the Isernhagen work systems functional capacity evaluation in patients with chronic low back pain. J Occup Rehabil. 2003;13:207–18.
26. Isernhagen SJ, Hart DL, Matheson LM. Reliability of independent observer judgments of level of lift effort in a kinesiophysical functional capacity evaluation. Work. 1999;12:145–50.
27. Gross DP, Battie MC. Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther. 2002;82:364–71.
28. WorkWell Systems Inc. Functional capacity evaluation V. 2nd ed. Duluth, MN: WorkWell Systems Inc; 2006.
29. Reneman MF, Kuijer W, Brouwer S, Preuper HR, Groothoff JW, Geertzen JH, et al. Symptom increase following a functional capacity evaluation in patients with chronic low back pain: an explorative study of safety. J Occup Rehabil. 2006;16:197–205.
30. Trippolini MA, Reneman MF, Jansen B, Dijkstra PU, Geertzen JH. Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. J Occup Rehabil. 2013;23:381–90.
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
32. Popping R. AGREE, a package for computing nominal scale agreement. Comput Stat Data Anal. 1984;2:182–5.
33. Randolph JJ. Online kappa calculator, 2008. Available from: http://justus.randolph.name/kappa. Accessed 21 June 2012.
34. Wind H, Gouttebarge V, Kuijer PP, Sluiter JK, Frings-Dresen MH. Effect of functional capacity evaluation information on the judgment of physicians about physical work ability in the context of disability claims. Int Arch Occup Environ Health. 2009;82:1087–96.
35. Gouttebarge V, Wind H, Kuijer PP, Sluiter JK, Frings-Dresen MH. Intra- and interrater reliability of the Ergo-Kit functional capacity evaluation method in adults without musculoskeletal complaints. Arch Phys Med Rehabil. 2005;86:2354–60.
36. Durand MJ, Loisel P, Poitras S, Mercier R, Stock SR, Lemaire J. The interrater reliability of a functional capacity evaluation: the physical work performance evaluation. J Occup Rehabil. 2004;14:119–29.
37. van Abbema R, Lakke SE, Reneman MF, van der Schans CP, van Haastert CJ, Geertzen JH, et al. Factors associated with functional capacity test results in patients with non-specific chronic low back pain: a systematic review. J Occup Rehabil. 2011;21:455–73.
38. Oesch P, Meyer K, Jansen B, Mowinckel P, Bachmann S, Hagen KB. What is the role of "nonorganic somatic components" in functional capacity evaluations in patients with chronic nonspecific low back pain undergoing fitness for work evaluation? Spine (Phila Pa 1976). 2012;37:E243–50.
39. Schapmire DW, St James JD, Townsend R, Feeler L. Accuracy of visual estimation in classifying effort during a lifting task. Work. 2011;40:445–57.
40. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43:543–9.
41. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. 4th ed. Oxford: Oxford University Press; 2008.
42. Moesch W. The use of objective parameters for functional capacity evaluations. Faculty of Medicine. Göttingen: Georg-August-Universität zu Göttingen; 2005. p. 70.
43. Verbunt JA, Seelen HA, Vlaeyen JW, Bousema EJ, Van Der Heijden GJ, Heuts PH, et al. Pain-related factors contributing to muscle inhibition in patients with chronic low back pain: an experimental investigation based on superimposed electrical stimulation. Clin J Pain. 2005;21:232–40.
44. Oesch P, Meyer K, Bachmann S, Hagen KB, Vollestad NK. Comparison of two methods for interpreting lifting performance during functional capacity evaluation. Phys Ther. 2012;92:1130–40.